Authorship recognition via fluctuation analysis of network topology and word intermittency

Diego R. Amancio

Department of Computer Science
Institute of Mathematical and Computer Sciences
University of São Paulo, São Carlos, São Paulo, Brazil

E-mail: `diego.raphael@gmail.com, diego@icmc.usp.br`

Abstract. Statistical methods have been widely employed in many practical natural language processing applications. More specifically, complex networks concepts and methods from dynamical systems theory have been successfully applied to recognize stylistic patterns in written texts. Despite the large amount of studies devoted to represent texts with physical models, only a few studies have assessed the relevance of attributes derived from the analysis of stylistic fluctuations. Because fluctuations represent a pivotal factor for characterizing a myriad of real systems, this study focused on the analysis of the properties of stylistic fluctuations in texts via topological analysis of complex networks and intermittency measurements. The results showed that different authors display distinct fluctuation patterns. In particular, it was found that it is possible to identify the authorship of books using the intermittency of specific words. Taken together, the results described here suggest that the patterns found in stylistic fluctuations could be used to analyze other related complex systems. Furthermore, the discovery of novel patterns related to textual stylistic fluctuations indicates that these patterns could be useful to improve the state of the art of many stylistic-based natural language processing tasks.

1. Introduction

The application of concepts from Physics in textual analysis has increasingly become widespread [1–7]. The use of entropy concepts is perhaps one of the most known examples of adapting methods from Physics in language-based models [8]. In recent years, physicists have proposed novel approaches to tackle several natural language processing problems [9–19]. The emergence of fundamental principles of organization common to all languages has been studied in terms of the least-effort principle [20]. Other studies have been devoted to the analysis of word frequency distributions [21–25], which has led to the design of novel cutting-edge keyword detection methods [16,28–32]. Syntactical features have been employed to investigate the fundamental properties of the language from a physical standpoint [16, 50, 51]. In the semantic/pragmatic level, concepts from Physics have also been used to investigate the ubiquity of ambiguous structures in texts [10, 26, 27].

In the field of stylometry, the use of complex networks (CN) in textual models has become commonplace [12, 15, 16, 33–35]. More specifically, several studies have modeled texts as co-occurrence (word adjacency) networks, where nodes and edges are represented by words and adjacency relationships, respectively. It has been shown that networks modeling texts share the same statistical properties of many other real systems [36]. Specially, such networks display both small-world and scale-free properties, as a consequence of the Zipf's law. Practical studies involving co-occurrence networks have devised algorithms to generate summaries [37], to assess text coherence and cohesion [38], and to evaluate the quality of manual and machine translations [39]. Even though word adjacency networks mostly grasp the syntactical factors of the language [16], it has been shown that they also convey semantic information [10, 26, 27].

While co-occurrence networks focus mainly on short scales, other physical models have been devised to capture long-range correlations. One of the most popular methods borrowed from the study of dynamical systems is the burstiness of word occurrences [29], which represents an attribute capable of capturing long-range textual features. Particularly, it has been shown that core words are unevenly distributed, while function words display distributions generated from random processes [31]. Such findings have motivated the proposition of algorithms aiming to detect keywords in single texts [30] using level statistics [33] and information theory [29]. The long-range textual structure has also been studied at the character unigram level [40, 41].

Most of the research on textual pattern recognition has focused on the search for recurring patterns in order to infer a specific class to unknown instances [42]. This approach has certainly worked well as many enlightening findings have been made this way. Despite the great number of studies on textual pattern recognition, the analysis of stylistic fluctuations along texts has received comparatively little attention. Empirical studies of some real systems have shown that fluctuations play a pivotal role on the unambiguous characterization of complex systems [43–47]. For example, when topology is a relevant network feature, the most informative patterns might be hidden in outlier fluctuations [48]. If one considers the distribution of word frequency, the fluctuations around the average might be useful to detect the most relevant concepts [29]. In biological systems, dynamical fluctuations of vital signals provide valuable information about the current state of the system [49].

Given the importance of the fluctuations in other real systems, the current paper presents a study on the properties of the stylistic variability along texts. Authors' styles were characterized upon measuring the topological connectivity of networks modeling texts [33]. The stylistic evolution was quantified upon splitting the texts in shorter subtexts, which in turn were represented as smaller networks. From the topological analysis of these varying networks, several interesting finding could be found. First and foremost, it was possible to identify the correct authorship of texts from a multivariate analysis of the stylistic fluctuations along literary works. Interestingly, in this model, the variability of the average shortest path lengths along subtexts turned out be the most relevant feature for discriminating distinct authors. To identify the authorship

of books, the proposed model also took advantage of the intermittency of time series representing the spatial distribution of words. Similarly to the CN-based model, a significant accuracy rate in discriminating authorship was found. Surprisingly, when the intermittency of 100 functional words was employed as features of the classifiers, the precise authorship could be found in 65% of the cases. As I shall show, the discovery of novel patterns related to the stylistic fluctuations in texts indicate that the proposed methodology can be extended to analyze other complex systems.

This paper is organized as follows. In Section 2, the methods employed to represent texts as networks are presented. This section also swiftly presents the main topological measurements employed for the characterization of complex networks. In the same section, the intermittency concept is presented. In Section 3, the authorship recognition task is studied. In this case, the variability of complex network measurements along texts and the intermittency of specific function words were employed as attributes of the classifiers for the authorship recognition task. Finally, Section 4 presents perspectives for further research.

2. Methods

In this paper, the style of written texts was quantified by measuring the topological properties of complex networks [60]. The representation of a text as a co-occurrence (word adjacency) network is detailed in Section 2.1. The topological features of complex networks employed to analyze the stylistic variation of texts are presented in Section 2.2. An alternative model based on the spatial distribution of words is presented in Section 2.3.

2.1. Modeling texts as complex networks

There are several ways to model texts as networks [52]. While semantic networks capture the relationships between word meanings, co-occurrence networks are more suitable to grasp stylistic attributes of written texts. As a matter of fact, co-occurrence networks represent a simplified version of syntactic networks [16] because most of the syntactic connections occurs between neighboring words [16, 56].

Prior to the creation of a co-occurrence network, some pre-processing steps are usually performed. Firstly, words conveying low semantic context (*stopwords*) are removed. Most of the words considered as *stopwords* are articles and prepositions (see the Supplementary Information). They are removed from the analysis because such words are mostly employed to connect other content words. After removing the *stopwords*, words with distinct spelling referring to the same concept are mapped to the same form. As a consequence, nouns and verbs are mapped to their singular and infinitive forms, respectively [53]. To perform such mapping, it is imperative to solve ambiguities at the word level because the mapped form might depend upon the sense assumed for a given word in a given context. To assist the disambiguation algorithm,

Table 1. Example of pre-processing steps performed to create a co-occurrence network. Firstly, *stopwords* are removed (see step #1). Then, the remaining words are converted to their canonical forms (see step #2). As a consequence, nouns and verbs are mapped to their singular and infinitive forms, respectively.

Original text	Step #1	Step #2
In the middle of the road	middle road	middle road
there was a stone there was	stone	stone
a stone in the middle of	stone middle	stone middle
the road there was a stone	road stone	road stone
in the middle of the road	middle road	middle road
there was a stone. Never	stone never	stone never
should I forget this event	I forget event	I forget event
in the life of my fatigued	life fatigued	life fatigue
retinas. Never should I	retinas never I	retina never I
forget that in the middle	forget middle	forget middle
of the road there was a	road	road
stone there was a stone	stone stone	stone stone
in the middle of the road	middle road	middle road
in the middle of the road	middle road	middle road
there was a stone.	stone	stone

the words are labeled with their respective parts-of-speech [53]. The labeling method employed is based on the maximum-entropy model proposed in [54].

After the pre-processing step, each distinct word becomes a node. Therefore, the total number of nodes in the network is equal to the vocabulary size (M) of the pre-processed text. The words that appear separated by up to $d - 1$ intermediate words are connected in the network. In this paper, the value $d = 1$ was used. Therefore, only adjacent words were connected. Table 1 illustrates the pre-processing steps taken to form a small network from the poem "In the middle of the road", by Carlos Drummond de Andrade. The network obtained from the pre-processed form is shown in Figure 1.

2.2. Topological characterization of complex networks

There are a myriad of measurements currently employed to characterize the topology of complex networks [55]. Traditional measurements can be classified according to the amount of information needed for the computation. While local measurements only require information about the neighbors of a given node, global measurements require that the global network connectivity is known beforehand. There is also a third class: the quasi-local measurements. As the name suggests, quasi-local measurements require information about further neighbors (i.e. the nodes located two or more hops away from the node under analysis). The following list swiftly describes the measurements employed to analyze the topology of networks modeling texts.
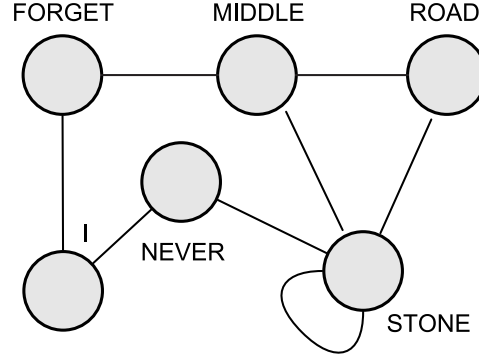
Figure 1. Example of co-occurrence network created for the poem "In the middle of the road", by Carlos Drummond de Andrade (see Table 1). Note that, after the removal of *stopwords*, adjacent words are connected (see first column of Table 1).

- *Clustering coefficient*: the clustering coefficient (C), a quasi-local measurement, quantifies the density of links between the neighbors of a given node. If c_i represents the number of edges between the neighbors of the node v_i and k_i is the total number of neighbors of v_i, the clustering coefficient is given by $C_i = 2c_i(k_i^2 - k_i)^{-1}$. In co-occurrence networks, the clustering coefficient measures the number of contexts in which a given word appears [33]. While generic words tend to take low values of clustering coefficient, context-specific words usually take higher values of clustering coefficient [33].

- *Average shortest path length*: to define this global measurement, consider we are given D_{ij}, the shortest distance between nodes v_i and v_j. The average shortest path length of v_i is then given by

$$l_i = \frac{1}{M(M-1)} \sum_i \sum_j D_{ij}. \tag{1}$$

In textual networks, this measurement quantifies the relevance of words. More specifically, a given word is considered relevant either when it is highly frequent or when it occurs close to the most relevant words [33].

- *Betweenness*: the betweenness (B) is a global measurement that measures the relevance of words [55]. To do so, the betweenness quantifies the number of shortest paths passing through a specific node. In textual networks, the betweenness also quantifies the number of contexts in which a word appear [33]. However, unlike the clustering coefficient, this measurement uses the global network information to infer the specificity of a word.

- *Accessibility*: this measurement is an extension of the degree k [37]. To define the accessibility, consider that $p_{ij}^{(h)}$ is the probability of a random walker to go from node v_i to node v_j in h steps. Mathematically, the accessibility (α) is computed from the irregularity (entropy) of the distribution of $p^{(h)}$

$$\alpha_i^{(h)} = \exp\left(-\sum p_{ij}^{(h)} \ln p_{ij}^{(h)}\right). \tag{2}$$

In general terms, the accessibility has been proven useful to identify the borders of complex networks when self-avoiding random walks are performed [37]. In textual networks, this measurement has been employed to identify keywords and to generate informative extractive summaries [37].

2.3. Intermittency

In linguistic models, the effects of attraction and repulsion of words is an ever present phenomenon [31, 57]. Several studies have shown that the distribution of many words along documents is not regular [29–32]. Particularly, keywords are usually unevenly distributed along texts [16, 28–32]. This finding has motivated the design of keyword detection methods relying upon a single document [53]. To analyze the spatial distribution of words, each token is mapped to a element in a temporal series. The first word of the text represents the first element, the second word represents the second element and so forth. Given a word w_i occurring f_i times in the text, the recurrence times of w_i generate the temporal series $T_i = \{t_1, t_2, t_3, \ldots, t_{f_i-1}\}$, where t_1 is the distance (i.e. the number of intermediary words) between the first and second occurrence of w_i, t_2 is the distance between the second and third occurrence of w_i and so on. Usually, two elements are added to the original temporal series T_i: the space t_0 until the first occurrence of w_i and the space t_{f_i} after the last occurrence of w_i. The distribution of T_i might be characterized by the mean and standard deviation:

$$\langle T \rangle_i = \frac{1}{f_i + 1} \sum_{i=0}^{f_i} t_i = \frac{N+1}{f_i + 1}, \tag{3}$$

$$\Delta T_i = \sqrt{\frac{1}{n_i} \sum_{i=0}^{n_i} (t_i - \langle T \rangle)}, \tag{4}$$

where $N = \sum f_i$. Given $\langle T \rangle$ and ΔT, the irregularity of the distribution T_i is computed as

$$I_i = \frac{\Delta T}{\langle T \rangle} = \frac{f_i + 1}{N+1} \sqrt{\frac{1}{f_i} \sum_{i=0}^{f_i} (t_i - \langle T \rangle)}. \tag{5}$$

The measurement defined in eq. 5 is known as intermittency (or burstiness) of the distribution. It has been widely employed to detect keywords in texts as an alternative to the tf-idf technique [53]. In addition, the intermittency has proven relevant to detect keywords in genetic sequences [32].

A qualitative comparison of words taking distinct values of intermittency is provided in Figure 2, which shows the distribution of the words "*Carmylle*" ($f_i = 54$) and "*feel*" ($f_i = 54$) along the book "Adventures of Sally", by Pelham Grenville Wodehouse. Because the distribution of "*Carmylle*" is much more irregular than the distribution of "*feel*", the former takes a much higher value of intermittency, as defined in eq. 5. The burstiness revealed by "*Carmylle*" also suggests that this word represents a relevant concept in the book [31]. An important property of the intermittency measurement

is that it does not correlate with the frequency (see Figure 3). This means that the relevance assigned by the intermittency is not influenced by the word frequency. Taking advantage of this property, recent studies have combined both frequency and intermittency measurements to improve several keyword detection methods [16, 30].
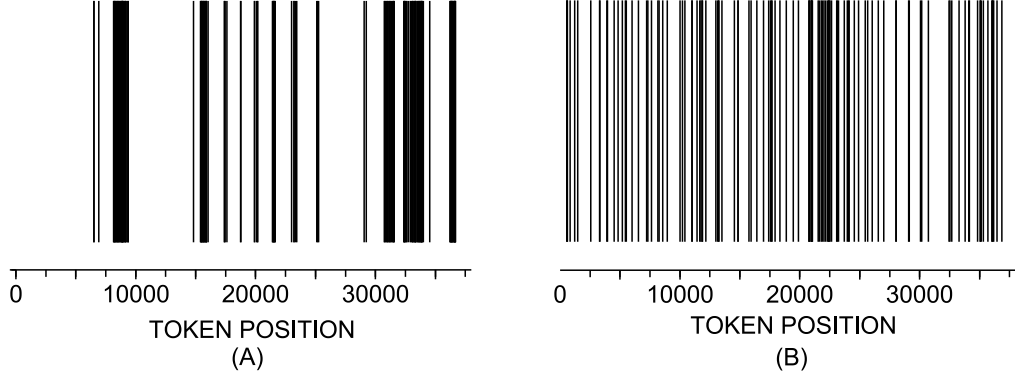

Figure 2. Profile of spatial distribution along the book "Adventures of Sally", by Pelham Grenville Wodehouse. The words considered were (a) "*Carmylle*"; and (b) "*feel*". Note that the distribution of "*Carmylle*" is much more irregular than the distribution of "*feel*". This suggests that "*Carmylle*" is much more relevant for the text than "*feel*".
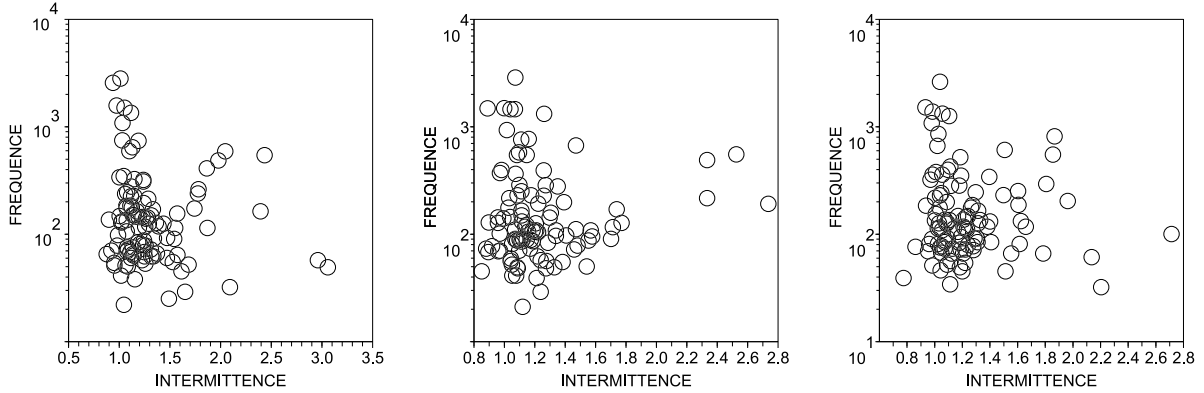

Figure 3. Pearson correlation coefficient between intermittency and frequency in the books (a) "Roughing It", by Mark Twain ($r = -0.14$); (b) "The woman in white", by Wilkie Collins ($r = 0.14$); and (c) "Moby Dick", by Herman Melville ($r = -0.17$).

2.4. Pattern recognition methods

The classification task aims at associating categories (or classes) to elements taking into account the attributes (or features) of these elements [65]. More specifically, an attribute is a measurable property of objects. To illustrate the concept, suppose that, in a given application, one desires to classify people according to their physical attributes. In this case, the height, the skin and hair color, the weight and others factors could

be selected as attributes. In many cases, the choice of discriminative and informative attributes plays an essential role on the performance of classification systems. Most of the attributes employed in traditional applications assume either numerical (e.g. 1, -7 and 3.14) or categorical values (e.g. *low* and *high*). In the current study, one of the attributes employed to characterize texts is the intermittency of specific words. In this case, the intermittency of each word represents a numeric attribute.

The classification task is of paramount relevance for information retrieval applications. Particularly, in this paper, pattern recognition methods are used to capture the patterns emerging from the representation of texts as networks. Moreover, pattern recognition methods are used to quantify the discriminative ability provided by these patterns. Currently, there are several automatic classification methods. They are traditionally divided into the following groups:

- **Supervised classification**: a binary relation mapping the input to the output is generated.

- **Unsupervised classification**: a partition of the dataset is generated so that similar elements are clustered together.

- **Semi-supervised classification**: the dataset available for automatic learning comprises a small set of labeled instances. Most of the instances is not labeled, i.e the class associated to these instances is lacking. In this case, the objective is to map the unlabeled input to a labeled output.

Typically, supervised classification methods process two datasets. The training dataset is the set of examples used as input. In other words, it represents the set of examples whose classes is known beforehand. In this paper, the training set is represented as $\mathcal{S}_{tr} = \{\beta_{(tr,1)}, \beta_{(tr,2)}, \beta_{(tr,3)} \ldots\}$. The test dataset $\mathcal{S}_{ts} = \{\beta_{(ts,1)}, \beta_{(ts,2)}, \beta_{(ts,3)} \ldots\}$ is the set used to evaluate the performance of the classifier. A given example β can be characterized by a set of $\mathcal{M}$ features: $\overrightarrow{\beta} = (F_1 = \beta^{(1)}, F_2 = \beta^{(2)}, \ldots, F_\mathcal{M} = \beta^{(\mathcal{M})})$, where $\mathcal{F} = \{F_1, F_2, \ldots, F_\mathcal{M}\}$ is the set of attributes characterizing the example β. In other words, the k-th value taken by the attribute F_k in β is represented as $\beta^{(k)}$. In a supervised classification, a given example assumes a single class c_i belonging to a finite set $\mathcal{C} = \{c_1, c_2, \ldots\}$.

To quantify the quality of the classification, the cross validation technique was employed [69]. In this method, a fraction of the dataset is used to perform the training and another fraction is used to perform the evaluation. The implementation of this technique consists in splitting the training dataset in ten folders. Initially, nine folders are selected to train the classifiers and the remaining one is used for evaluation. This process is repeated ten times so that a different folder is used for the evaluation in each iteration. Finally, the accuracy rate is computed as the average accuracy obtained over the ten iterations. The cross validation is considered a reliable index since the evaluation is performed over unknown instances.

In the experiments, the analysis was performed on a dataset comprising books whose authorship is known beforehand. As a consequence, supervised classification methods

were employed to recognize patterns in the generated textual time series. The methods employed in this study were: Bayesian Networks (BNT), Complement Naive Bayes (CNB), Naive Bayes (NVB), RBF Networks (RBF), Multi Layer Perceptron (MLP), Support Vector Machines (SVM), k Nearest Neighbors (KNN), C4.5 (C45) and Random Forest (RFO). A short introduction to these methods is provided in Appendix A.

3. Results

The stylistic properties of texts was studied in the context of the authorship recognition task. In this problem, one tries to recognize the identity of authors whose authorship is unknown. Owing to its central importance for stylometry, several contributions have been proposed [58]. Simple approaches include the analysis of word length and additional character features [67]. Mosteller and Wallace [68] proved that the frequency of function words (such as "*and*", "*any*", "*ever*", "*or*", "*until*" and "*with*") can be employed to quantify the style of authors. More recently, many other approaches have been devised [58], including those relying upon topological analysis of complex networks [33, 70]. Here I use complex network and intermittency measurements to obtain potentially useful attributes for identifying the authors of books whose identity is lacking. Because this study focus on the analysis of stylistic fluctuations, the patterns displayed by the evolution of the statistical measurements along texts were studied.

Two types of temporal series representing the stylistic evolution in books are studied. In Section 3.1, the relationship between the stylistic variation along books and the authorship recognition task is investigated. In Section 3.2, the intermittency profile of some words across different authors is employed to perform the authorship recognition task. The dataset employed in the experiments comprises books written by 8 authors, as shown in Table S1 of the Supplementary Information.

3.1. Stylistic variation along books

In this section, I investigate whether the stylistic variation along texts provides useful attributes to the authorship recognition task. To quantify stylistic variations, the following methodology was taken. Each book in the dataset was split in subtexts comprising W tokens. Assuming that a book is formed by a sequence of tokens $\mathcal{W} = \{w_1, w_2, \ldots\}$, the j-th subtext $\mathcal{T}_j$ will comprise the sequence $\{w_{S_j}, w_{S_j+1}, \ldots, w_{S_j+W}\}$, where $S_j = W \cdot j + 1$ and $j \in \{0, 1, 2, \ldots\}$. Each subtext $\mathcal{T}_i$ was modeled as a complex network (see Section 2.1) and the topological measurements of each subtext were extracted (see Section 2.2). Thus, each topological measurement X generates a temporal series $\mathcal{X} = \{x_1, x_2, \ldots x_P\}$, where x_i represents the value obtained for X in the subtext $\mathcal{T}_i$ and P is the total number of subtexts. An example of $\mathcal{X}$ for $X = \langle l \rangle$ and $X = \langle C \rangle$ is provided in Figure 4. The temporal series $\mathcal{X}$ of each book was then

decomposed in terms of the Fourier transform:

$$\mathscr{F}(X)_{(j)} = \sum_{k=1}^{P} x_k \exp\left(\frac{-2\pi ijk}{N}\right),$$ (6)

where $i^2 = -1$. It is worth noting that the first component, given by

$$\mathscr{F}(X)_{(0)} = \sum_{k=1}^{P} x_j = P\langle x\rangle,$$

only stores information concerning the average $\langle x\rangle$. Higher frequencies and, therefore, higher levels of variation in $\mathcal{X}$ are represented in $\mathscr{F}(X)_{(\{j\in\mathcal{N}|j\geq 1\})}$. As attributes of the classifiers, the first four components of $\mathscr{F}(X)$ were used.
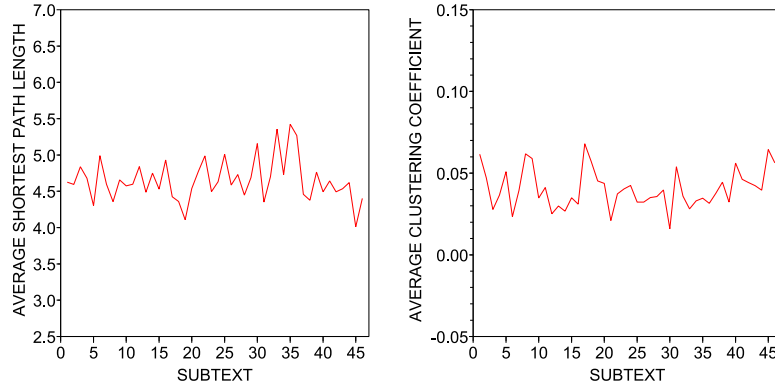

Figure 4. Example of topological variation along the book "Great Expectation", by Charles Dickens. The networks were formed using $W = 1,300$ tokens. The measurements considered were (a) the average shortest path length $\langle l\rangle$; and (b) the average clustering coefficient $\langle C\rangle$.

The results obtained from the classification of authors are shown in Table 2. The length of the subtexts considered were $W = \{500, 700, 900, 1,100, 1,300\}$. For each subtext length, the table lists the accuracy rate obtained by the best classifier. The lowest accuracy rate occurred for $W = 500$ and the highest discriminability was achieved with $W = 1,300$. In all cases, the performance obtained by the classifiers was statistically significant, as revealed by low p-values. This result confirms that the stylistic variations of authors along texts (quantified via topological analysis of complex networks) can be employed to discriminate authors' styles. Specially, the proposed method could be used as a complementary stylistic attribute, because the stylistic variation has been widely neglected as a relevant feature in current authorship attribution methods [58].

To verify the relative relevance of the features employed in the authorship recognition task, the information gain of each attribute in the training dataset was computed. Mathematically, the relevance ascribed by the information gain (Ω) is

$$\Omega(\mathcal{S}_{tr}, F_k) = \mathcal{H}(\mathcal{S}_{tr}) - \mathcal{H}(\mathcal{S}_{tr}|F_k),$$ (7)

Table 2. Accuracy rate obtained in the classification based on the Fourier decomposition of time series of complex networks measurements. For each subtext length (W), the table lists the accuracy rate obtained by the best classifier.

W	Method	Accuracy	p-value
500	BNT	35.0%	2.2×10^{-4}
700	CNB	37.5%	5.2×10^{-5}
900	CNB	40.0%	1.1×10^{-5}
1,100	RBF	42.5%	2.2×10^{-6}
1,300	RFO	45.0%	4.0×10^{-7}

where $\mathcal{H}(\mathcal{S}_{tr})$ is the entropy of the training dataset $\mathcal{S}_{tr}$ and $\mathcal{H}(\mathcal{S}_{tr}|F_k)$ is the entropy of $\mathcal{S}_{tr}$ when F_k is specified. $\mathcal{H}(\mathcal{S}_{tr}|F_k)$ can be computed from training dataset as

$$\mathcal{H}(\mathcal{S}_{tr}|F_k) = \sum_{v \in V(F_k)} |\beta_{(tr)}^{(k)} = v| \cdot |\mathcal{S}_{tr}|^{-1} \cdot \mathcal{H}(\beta_{(tr)}^{(k)} = v), \tag{8}$$

where $|\cdot|$ is the cardinality of the set and $V(F_k)$ represents the set of all values taken by the attribute F_k in the training dataset, i.e.

$$V(F_k) = \bigcup_{i=1}^{|\mathcal{S}_{tr}|} \beta_{(tr,i)}^{(k)}. \tag{9}$$

The rank of the most informative measurements, according to eq. 7, is shown in Table 3. In this table, the rows indicate the ranking obtained by the attributes, for each subtext length (W). All in all, the vocabulary size M turned out to be one of the most relevant attributes. The measurement displaying the highest relevance in higher components of the Fourier transform ($j \geq 1$ in eq. 6) was the average shortest path length. More specifically, the third component ($j = 3$) displayed the highest relevance for large values of W, suggesting that the attribute related to the variation of $\langle l \rangle$ along the text becomes even more relevant when larger subtexts are analyzed. Interestingly, this result reinforces the importance of shortest paths for the authorship attribution task, since this measurement has been successfully employed to characterize authors' styles in networks formed from full books [33]. The relevance of higher components of the Fourier transform can also be noted in the decision tree built with the C4.5 method [59] (see Figure 5). Note that $\mathscr{F}(\langle l \rangle)_{(2)}$ and $\mathscr{F}(M)_{(2)}$ appear at superior levels of the tree, confirming thus their relevance. The relative importance of higher components becomes even more apparent if one observes that some traditional complex network measurements (i.e. $\mathscr{F}(X)_{(0)}$) correlates with the vocabulary size M [33]. This does not occur with $\mathscr{F}(\langle l \rangle)_{(2)}$, as revealed by the Pearson correlation coefficient displayed in Table 4. In fact, none of the higher components found in Table 4 correlates significantly with other relevant traditional attributes ($\mathscr{F}(X)_{(0)}$), thus confirming that higher components indeed provide novel information for characterizing styles in written texts.

The results concerning the evolution of styles revealed that distinct authors might display distinct stylistic patterns along texts. This finding is similar to the results found

Table 3. Relative importance of the attributes used in the classification based on the spectral decomposition of complex network measurements. The rows represent the ranking obtained for a given attribute. For example, the best attribute for $W = 1,300$ was $\mathscr{F}(\langle C \rangle)_{(0)}$ and the second best attribute for $W = 1,300$ was $\mathscr{F}(\langle M \rangle)_{(0)}$. The measurements taking values of information gain below 0.500 are not shown. According to the information gain index, the third component of the average shortest paths lengths turned out to be one of the most informative measurement.

#	W=500	W=700	W=900	W=1,100	W=1,300
1st	$\mathscr{F}(M)_{(0)}$ 0.772	$\mathscr{F}(M)_{(0)}$ 0.812	$\mathscr{F}(\langle C \rangle)_{(0)}$ 0.778	$\mathscr{F}(\langle l \rangle)_{(2)}$ 0.850	$\mathscr{F}(\langle C \rangle)_{(0)}$ 0.778
2nd	$\mathscr{F}(\langle \alpha^{(3)} \rangle)_{(0)}$ 0.669	$\mathscr{F}(\langle C \rangle)_{(0)}$ 0.778	$\mathscr{F}(M)_{(0)}$ 0.772	$\mathscr{F}(M)_{(0)}$ 0.772	$\mathscr{F}(M)_{(0)}$ 0.772
3rd	$\mathscr{F}(\langle l \rangle)_{(0)}$ 0.665	$\mathscr{F}(\langle l \rangle)_{(0)}$ 0.772	$\mathscr{F}(\langle l \rangle)_{(0)}$ 0.712	$\mathscr{F}(\langle l \rangle)_{(0)}$ 0.691	$\mathscr{F}(\langle l \rangle)_{(0)}$ 0.712
4th	$\mathscr{F}(\langle \alpha^{(2)} \rangle)_{(0)}$ 0.653	$\mathscr{F}(\langle \alpha^{(3)} \rangle)_{(0)}$ 0.669	$\mathscr{F}(\langle \alpha^{(3)} \rangle)_{(0)}$ 0.653	$\mathscr{F}(\langle \alpha^{(3)} \rangle)_{(0)}$ 0.601	$\mathscr{F}(\langle l \rangle)_{(2)}$ 0.608
5th	$\mathscr{F}(\langle l \rangle)_{(2)}$ 0.558	$\mathscr{F}(\langle \alpha^{(2)} \rangle)_{(0)}$ 0.669	$\mathscr{F}(\langle \alpha^{(2)} \rangle)_{(0)}$ 0.653	$\mathscr{F}(\langle \alpha^{(2)} \rangle)_{(0)}$ 0.601	$\mathscr{F}(\langle \alpha^{(3)} \rangle)_{(0)}$ 0.606
6th		$\mathscr{F}(\langle l \rangle)_{(2)}$ 0.558	$\mathscr{F}(\langle l \rangle)_{(2)}$ 0.548		$\mathscr{F}(\langle \alpha^{(2)} \rangle)_{(0)}$ 0.601
7th					$\mathscr{F}(\langle M \rangle)_{(2)}$ 0.558
8th					$\mathscr{F}(\langle \alpha^{(3)} \rangle)_{(2)}$ 0.510

in [60], which showed that the temporal evolution of stylistic features of books published between 1590 and 1922 is able to identify the traditional literary movements. The main feature differentiating this work from previous studies is that the stylistic variation inside books is much more subtle than the corresponding variation over different literary styles [61]. The emergence of the described patterns suggests the applicability of other temporal models. Alternative models could probe, for example, the patterns present in the spatial distribution of character bigrams [40]. Particularly, this paper focus primarily on the evolution of stylistic patterns measured by the spatial distribution of words. For this reason, the next section investigates if the intermittency of specific words serves as authors' fingerprints for the authorship recognition task.

3.2. Authorship recognition via word intermittency

To verify if the uneven distribution of specific words along texts provides useful features for characterizing authors' styles, the following experiment was carried out. Following the research on stylometry, this study focused on function words. The 100 most frequent
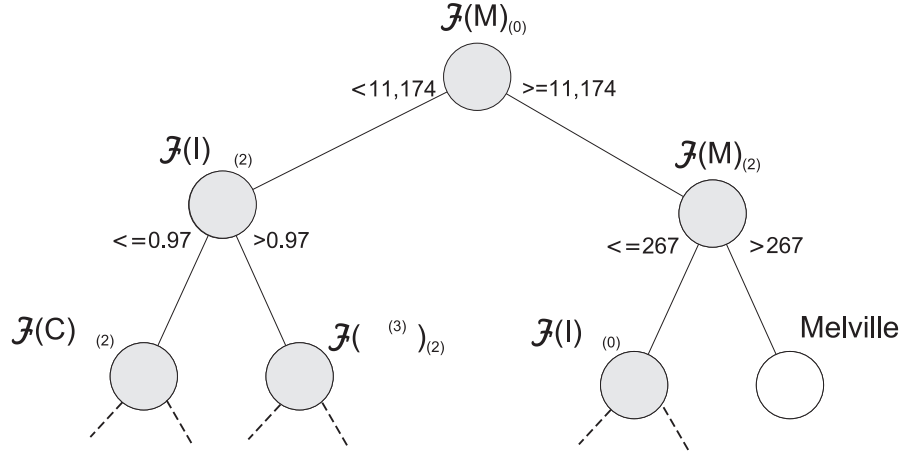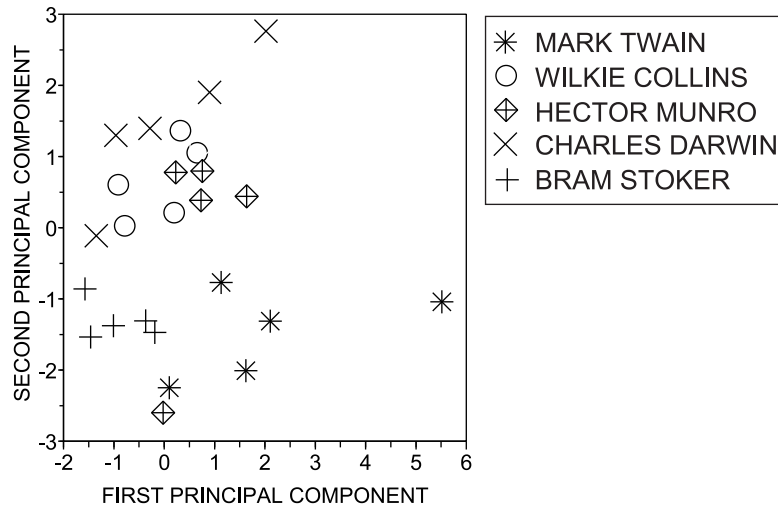
Figure 5. Example of decision tree created to identify the authorship of books. To construct the tree, the C4.5 algorithm was employed in subtexts comprising $W = 1,300$ tokens. Note that the second component of the average shortest path length and vocabulary size ($\mathscr{F}(\langle l \rangle)_{(2)}$ and $\mathscr{F}(\langle M \rangle)_{(2)}$) are relevant as they appear at the top of the tree.

Table 4. Pearson correlation coefficient $|r|$ between $\mathscr{F}(X)_{(j=0)}$ and the most informative measurements found for $W = 1,300$ (see Table 3). Because all correlations assume low values, the information conveyed by $\mathscr{F}(X)_{(2)}$ differs from the simple average $\langle X \rangle = \mathscr{F}(X)_{(0)}$.

x	y	$|r(x,y)|$
$\mathscr{F}(\langle l \rangle)_{(2)}$	$\mathscr{F}(\langle C \rangle)_{(0)}$	0.073
$\mathscr{F}(\langle l \rangle)_{(2)}$	$\mathscr{F}(M)_{(0)}$	0.182
$\mathscr{F}(\langle l \rangle)_{(2)}$	$\mathscr{F}(\langle l \rangle)_{(0)}$	0.170
$\mathscr{F}(\langle l \rangle)_{(2)}$	$\mathscr{F}(\langle \alpha^{(3)} \rangle)_{(0)}$	0.015
$\mathscr{F}(\langle l \rangle)_{(2)}$	$\mathscr{F}(\langle \alpha^{(2)} \rangle)_{(0)}$	0.100
$\mathscr{F}(M)_{(2)}$	$\mathscr{F}(\langle C \rangle)_{(0)}$	0.209
$\mathscr{F}(M)_{(2)}$	$\mathscr{F}(M)_{(0)}$	0.041
$\mathscr{F}(M)_{(2)}$	$\mathscr{F}(\langle l \rangle)_{(0)}$	0.036
$\mathscr{F}(M)_{(2)}$	$\mathscr{F}(\langle \alpha^{(3)} \rangle)_{(0)}$	0.059
$\mathscr{F}(M)_{(2)}$	$\mathscr{F}(\langle \alpha^{(2)} \rangle)_{(0)}$	0.115
$\mathscr{F}(\langle \alpha^{(3)} \rangle)_{(2)}$	$\mathscr{F}(\langle C \rangle)_{(0)}$	0.010
$\mathscr{F}(\langle \alpha^{(3)} \rangle)_{(2)}$	$\mathscr{F}(M)_{(0)}$	0.042
$\mathscr{F}(\langle \alpha^{(3)} \rangle)_{(2)}$	$\mathscr{F}(\langle l \rangle)_{(0)}$	0.045
$\mathscr{F}(\langle \alpha^{(3)} \rangle)_{(2)}$	$\mathscr{F}(\langle \alpha^{(3)} \rangle)_{(0)}$	0.080
$\mathscr{F}(\langle \alpha^{(3)} \rangle)_{(2)}$	$\mathscr{F}(\langle \alpha^{(2)} \rangle)_{(0)}$	0.077

words in the corpus were considered as function words. As such, as attributes for the classifiers, the intermittency of these function words was used. The best classifier, the Multilayer Perceptron, yielded an accuracy rate of 65.0% (p-value $= 1.3 \times 10^{-14}$). This result suggests that, besides the frequency, the intermittency of specific function words might be useful for characterizing authors' styles in texts. Note that the discriminability obtained with intermittency features is not influenced by the frequency of function words, since there is no significant correlation between intermittency and frequency (see Section 2.3).

A detailed analysis of the classification revealed that most of the errors occurred for Arthur Conan Doyle, Wilkie Collins and Mark Twain (result not shown). If these authors are disregarded from the analyis, the use of intermittency features would provide an accuracy rate of 90% with the Multilayer Perceptron. Despite the large number of attributes employed for discriminating authors' styles, the discriminative ability concentrated in a few function words. According to the information gain measurement, the words displaying the highest discriminative ability were "*but*" ($\mathcal{H} = 0.620$), "*and*" ($\mathcal{H} = 0.604$), "*I*" ($\mathcal{H} = 0.530$), "*who*" ($\mathcal{H} = 0.494$) and "*as*" ($\mathcal{H} = 0.462$). The high discriminability obtained with the intermittency of these five words can be noted in the principal component analysis shown in Figure 6.

Figure 6. Principal component analysis performed for the authorship recognition task. As attributes, only the five most informative features (the intermittency of "*but*", "*and*", "*I*", "*who*" and "*as*") were employed to create the figure.

In summary, one can conclude that the representation of specific words as temporal series might be useful for the authorship recognition task. As commented in Section 3.1, the use of intermittency of specific words combined with traditional features might be useful to improve the performance of style-based real applications. In this case, an improved textual characterization would be provided, because the attributes generated from textual fluctuations do not correlate with traditional features. Moreover, distinct classifiers could be employed for each attribute type (e.g. frequency or intermittency), as

some classifiers perform better for specific attributes. As such, the classification becomes more robust and accurate without the fine tuning required in single models [71]. The combination of attributes could be performed via ordinary voting of simple models [72]. Another possibility is to consider fuzzy methods as independent classifiers and then select the best weighting strategy for each classifier [26]. Furthermore, the successful application of intermittency measurements in characterizing authors' styles suggests that complementary studies should be carried out in order to probe whether additional features of temporal series modeling the spatial distribution of words are able to reveal novel stylistic/topological patterns.

4. Conclusion

In this study, I investigated if measurements characterizing temporal series from texts are useful to identify authors' styles. In the light of the results, one can conclude that authors' stylistic properties can be characterized upon analyzing the fluctuations of textual statistical measurements. The statistically significant accuracy rates obtained in the authorship attribution task confirmed that the features derived from the fluctuation of specific topological and intermittency measurements are able to discriminate distinct authors. Using a co-occurrence network model, it was shown that the relative importance of distinct attributes may depend on the subtext length. Nevertheless, in general, further components of the Fourier decomposition of topological measurements turned out to be relevant features for the task. An analysis of the spatial distribution of specific words revealed distinct patterns of distribution for different authors. Surprisingly, the intermittence of functional words correctly discriminate the authorship in 65% of the cases in a dataset comprising books written by 8 authors.

The focus of this investigation was on the evaluation of distinct attributes for characterizing authors' styles, rather than maximizing the accuracy rate of the classification. However, the dependence with stylistic attributes found for the proposed features suggests that attributes derived from the analysis of stylistic fluctuations can be combined in a hybrid way with traditional attributes, such as the frequency of function words [68]. As such, the findings reported in this paper shall potentially contribute to the improvement of current authorship recognition methods [67]. One could pursue this line of analysis further, identifying the combination of features yielding the best discriminability. Future investigations could probe the relevance of fluctuations in other related complex systems, such as DNA and other generic symbolic sequences, since the techniques described here can be extended in a straightforward fashion to such cases.

Acknowledgments

Appendix A. Pattern Recognition Methods

This appendix swiftly describes the main pattern recognition methods employed in this study. A complete reference to the field of pattern recognition can be found in [66].

Decision trees

Decision tree algorithms employ trees [62] to summarize the patterns recognized in the dataset (see Figure A1). Typically, a decision tree comprises internal and leaf nodes. While internal nodes store the tests performed on specific attributes, leaf nodes represent classes. The edges connect nodes according to the answers obtained from the tests. For example, the node representing the test $F_1 > \theta_1$ has two outgoing edges, namely "YES" and "NO". During the classification stage, one travels through the tree until a leaf node is reached. In this case, the class associated to the leaf node is assigned to the unknown instance. The classification process is illustrated in Figure A1.
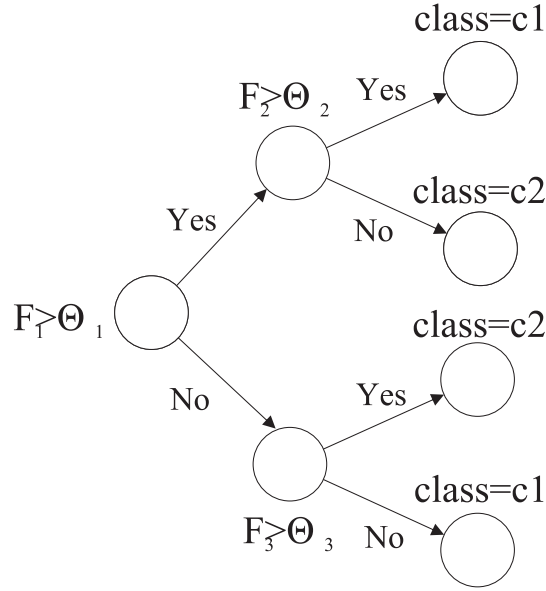

Figure A1. Example of decision tree. To classify a new instance, one starts the walk at the root node. The class assigned to the unknown instance is the class associated to the leaf node found at the end of the walk.

To construct a decision tree, at each step, one tries to find an attribute F_i and a threshold θ so that the test $F_i \geq \theta$ yields the best dataset partition. One assumes that the quality of a partition is proportional to the discriminability provided by that partition. At each division, the goal is to separate one or more classes in distinct groups. Several measurements have been proposed to quantify the quality of partitions. An well-known measurement is the Kullback-Leibler divergence [63]. The process of choosing the attribute with the highest information gain is reiterated for the two subsets created at each internal node. The recursion is finalized when a subset contains instances belonging to a single class. In this case, a leaf node is created to store the corresponding class.

The tree-based algorithms employed in this paper were the C4.5 and Random Forest. Further details regarding these methods can be found in [69].

Bayesian decision

To classify a new instance, the Naive Bayes algorithm estimates the probability distribution of each class $c_i \in \mathcal{C}$. Given the likelihood profile of each class, the algorithm employs the *maximum a posteriori* strategy to infer the correct class. The probability of each $c_i \in C$ to be assigned to the instance β is

$$P(c_i|\overrightarrow{\beta}) = \frac{P(\overrightarrow{\beta}|c_i)P(c_i)}{P(\overrightarrow{\beta})}$$

$$= \frac{P(F_1 = \beta^{(1)}, \dots, F_{\mathcal{M}} = \beta^{(\mathcal{M})}|c_i)P(c_i)}{P(F_1 = \beta^{(1)}, \dots, F_{\mathcal{M}} = \beta^{(\mathcal{M})})}.$$

Note that $P(c_i)$ can be estimated as $\mathcal{N}(c_i)/\sum_{c_i \in \mathcal{C}} \mathcal{N}(c_i)$, where $\mathcal{N}(c_i)$ is the number of objects in $\mathcal{S}_{tr}$ belonging to class c_i. For classification purposes, the quantity $P(\overrightarrow{\beta})$ can be disregarded from the analysis because $P(\overrightarrow{\beta}|c_i)$ is constant for all $c_i \in \mathcal{C}$. Finally, in order to estimate $P(\overrightarrow{\beta}|c_i)$, the traditional Naive Bayes classifier surmises independence between the features. Hence $P(\overrightarrow{\beta}|c_i)$ is estimated as

$$P(\overrightarrow{\beta}|c_i) = P(F_1 = \beta^{(1)}, \dots, F_{\mathcal{M}} = \beta^{(\mathcal{M})}|c_i)$$

$$= \prod_{k=1}^{\mathcal{M}} P(F_k = \beta^{(k)}|c_i).$$

Using the value of $P(\overrightarrow{\beta}|c_i)$, it is possible to replace it in the definition of $P(c_i|\overrightarrow{\beta})$. Therefore

$$P(c_i|\overrightarrow{\beta}) = \frac{P(c_i)}{P(F_1 = \beta^{(1)}, \dots)} \prod_{k=1}^{\mathcal{M}} P(F_k = \beta^{(\mathcal{M})}|c_i).$$

Upon using the *maximum a posteriori* rule, the class c_s can be estimated as

$$c_\beta = \arg\max_{c_i \in \mathcal{C}} P(c_i) \prod_{k=1}^{\mathcal{M}} P(F_k = \beta^{(k)}|c_i).$$

To obtain c_β from the above equation, one must estimate the likelihood $P(F_k|c_i)$. Several methods have been proposed to perform the estimation [64]. The Parzen-Rosenblatt window algorithm has been widely employed as a non-parametric technique to estimate probability densities [64].

In addition to the Naive Bayes, the algorithms based on statistical paradigms employed in this study were the Complement Naive Bayes and Bayesian Networks. More details concerning these methods can be found in [69].

Neural Networks

The simplest artificial neural network (ANN) model is the Perceptron. In this model, each neuron stores activation and transfer functions. While the former sums (with
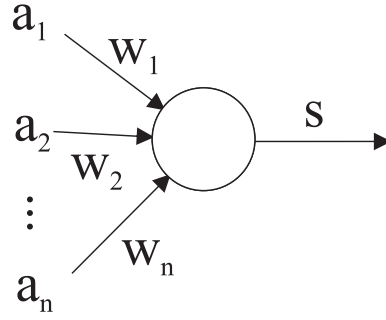
Figure A2. Example of a single neuron. We are given some input signals a_i and expected outputs in a supervised classification. The learning algorithm aims at minimizing the error between the actual and expected output signals.

weights) the input signals, the latter yields an output signal as a function of the input. Figure A2 illustrates a single neuron with input signals and weights represented as a_i and w_i, respectively. The output s is $s = \sum_i a_i w_i + b$. The transfer function ϕ may assume many distinct forms [65]. A very simple possibility is to consider that the neuron is activated whenever s surpasses a given threshold, i.e.

$$\phi(s) = \begin{cases} 1 & \text{for } s > 0, \\ 0 & \text{otherwise.} \end{cases} \tag{A.1}$$

The correct choice of synaptic weights in neural networks allows the network to effectively process the input signals in order to generate the expected output. In general, the weights are assigned by learning algorithms [65]. Initially, the values w_{ij} of weights linking the i-th node of the input layer with the j-th node of the output layer assume random values. Given these initial weights, several input signals are presented to the neuron. Then, the obtained output is compared with the expected values. If the observed error exceeds a given threshold, the current weights are modified by the learning algorithm. In this case, the larger the error obtained, the greater is the change applied to the current weights. More specifically, weights are updated according to the rule $w_{ij}^{(t+1)} = w_i^{(t)} + \eta \varepsilon_j x_i$, where η is the learning rate and ε_j is the error obtained for the j-th neuron.

The ANN-based pattern recognition methods employed in this study were the Multilayer Perceptron and the RBF network. More details concerning these methods can be found in [65].

References

[1] Pop C-M and Frey E 2013, Phys. Rev. E **88** 022814

[2] Ausloos M 2012 Phys. Rev. E **86** 031108

[3] Martinez-Romo J, Araujo L, Borge-Holthoefer J, Arenas A, Capitn J and Cuesta J 2011 Phys. Rev. E. **84** 46108

[4] Mehri A and Darooneh A H 2011 Physica A **390** 3157–3163

[5] Isern N and Fort J 2014 J. R. Soc. Interface **11** 94

[6] Font-Clos F, Boleda G and Corral A 2013 New J. Phys. **15** 093033

[7] Huang J 2013 PLoS ONE **8** e74515

[8] Berger A L, Della Pietra V J and Della Pietra S A 1996 Comput. Linguist. **22**, 39–71

[9] Sienkiewicz J, Skowron M, Paltoglou G and Yst J A H 2013 Adv. Complex Syst. **16** 1350026

[10] Amancio D R, Oliveira Jr. O N and Costa L F 2012 Europhys. Lett. **98** 18002

[11] Yang Z, Lei J, Fan K and Lai Y 2013 Physica A **392** 4523–4531

[12] Amancio D R, Oliveira Jr. O N and Costa L F 2012 Physica A **391** 4406–4419

[13] Ausloos M 2012 Chaos Soliton Fract. **45** 1349–1357

[14] Yasseri T, Kornai A and Kertsz J 2012 PLoS ONE **7** e48386

[15] Amancio D R, Aluisio S M, Oliveira Jr. O N and Costa L F 2012 Europhys. Lett. **100** 58002

[16] Amancio D R, Altmann E G, Rybski D, Oliveira Jr. O N and Costa L F 2013 PLoS ONE **8** e67310

[17] Amancio D R, Oliveira Jr. O N and Costa L F 2012 J. Stat. Mech. Theor. Exp. P01004

[18] Li W 2012 Physica A **391** 1515–1518

[19] Holanda A J, Pisa I T, Kinouchi O, Martinez A S and Ruiz E E S 2004 Physica A **344** 530–536

[20] Ferrer-i-Cancho R and Solé R V 2003 Proc. Natl. Acad. Sci. USA **100** 788–791

[21] Allahverdyan A E, Deng W and Wang Q A 2013 Phys. Rev. E **88** 062804

[22] Baixeries J, Elvevag B and Ferrer-i-Cancho R 2013 PLoS ONE **8** e53227

[23] Corominas-Murtra B, Fortuny J and Solé R V 2011 Phys. Rev. E **83** 036115

[24] Eliazar I 2011 Physica A **390** 3189–3203

[25] Ferrer-i-Cancho R and Elvevag B 2010 PLoS ONE **5** e9411

[26] Silva T C and Amancio D R 2012 Europhys. Lett. **98** 58001

[27] Montemurro M A 2014 Cortex **55C** 5–16

[28] Mathiesen J, Angheluta L and Jensen M H 2014 Eur. Phys. J-Spec. Top. **223** 1849–1858

[29] Carretero-Campos C, Bernaola-Galvan P, Coronado A V and Carpena P 2013 Physica A **392** 1481–1492

[30] Carpena P, Bernaola-Galvan P, Hackenberg M, Coronado A and Oliver J 2009 Phys. Rev. E **79** 035102(R)

[31] Herrera J P and Pury P A 2008 Eur. Phys. J. B 63 135–146

[32] Ortuno M, Carpena P, Bernaola-Galvan P, Munoz E and Somoza A M 2002 Europhys. Lett. **57** 759–764

[33] Amancio D R, Altmann E G, Oliveira Jr. O N and Costa L F 2011 New J. Phys. **13** 123024

[34] Sheng L and Li C 2009 Physica A **388** 2561–2570

[35] Xuan Q, Wu T-J 2009 Phys. Rev. E **80** 026103

[36] Ferrer-i-Cancho R and Solé R V 2001 Proc. R. Soc. B **268** 2261–2266

[37] Amancio D R, Nunes M G V, Oliveira Jr. O N and Costa L F 2012 Physica A **391**, 1855–1864

[38] Antiqueira L, Nunes M G V, Oliveira Jr. O N and Costa L F 2007 Physica A **373** 811–820

[39] Amancio D R, Antiqueira L, Pardo T A S, Costa L F, Oliveira Jr. O N and Nunes M G V 2008 Int. J. Mod. Phys. C **19** 583–598

[40] Escalante H J, Solorio T and Montes-y-Gomez M 2011 *Proceeding HLT '11 Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies* **1** 288–298

[41] Tomovic A, Janicic P and Keselj V 2006 Comput. Meth. Prog. Bio. **81** 137–153

[42] Kudo M and Sklansky J 2000 Pattern Recognition **33** 25–41

[43] Eisler Z, Bartos I and Kertesz J 2008 Adv. Phys. **57** 89–142

[44] Medina J M, Diaz J A and Norwich K H 2014 Front. Hum. Neurosci. **8**

[45] Kalyuzhny M, Schreiber Y, Chocron R, Flather C H, Kadmon R, Kessler D A and Shnerb N M 2014 Ecology 95 1701–1709

[46] Chen Q, Qian J-H and Han D-D 2014 Int. J. Mod. Phys. C **25** 1440012

[47] Blumm N, Ghoshal G, Forró Z, Schich M, Bianconi G, Bouchaud J-P and Barabási A-L 2012 Phys. Rev. Lett. **109** 128701

[48] Costa L F, Rodrigues F A, Hilgetag C C and Kaiser M 2009 Europhys. Lett. **87** 18008

[49] Sato K, Ito Y, Yomo T and Kaneko K 2003 P. Natl. Acad. Sci. USA **100** 14086–14090

[50] Liu H 2008 Physica A **387** 3048–3058

[51] Liu H and Hu F 2008 Europhys. Lett. **83** 18002

[52] Solé R V, Corominas-Murtra B, Valverde S and Steels L 2010 Complexity **15** 20–26

[53] Manning C D and Schutze H 1999 *Foundations of Statistical Natural Language Processing*, Cambridge, MA: MIT Press

[54] Kohavi R 1995 *Proceedings of the International Joint Conference on Artificial Intelligence* **2** 12

[55] Newman M 2010 *Networks: An Introduction*, Oxford University Press, Inc., New York, USA

[56] Ferrer-i-Cancho R, Solé R V and Kohler R 2004 Phys. Rev. E **69** 051915

[57] Bussemaker H J, Li H and Siggia E D 2000 Proc. Natl. Acad. Sci. USA **97** 10096

[58] Stamatatos E 2009 J. Am. Soc. Inf. Sci. Technol. **60** 538–556

[59] Quinlan J R 1993 C4.5: Programs for Machine Learning, Morgan Kaufmann Publishers

[60] Amancio D R, Oliveira Jr. O N and Costa L F 2012 New J. Phys. **14** 043029

[61] Amancio D R 2014 *Probing the topological properties of complex networks modeling short written texts.* Manuscript under review

[62] Cormen T H, Leiserson C E, Rivest R L and Stein C 2001 *Introduction to Algorithms*, MIT Press and McGraw-Hill

[63] Kullback S and Leibler R A 1951 Ann. Math. Stat. **22** 79–86

[64] Lozano J A, Larranga P, Inza I and Bengoetxea E 2006 *Towards a new evolutionary computation. Advances in estimation of distribution algorithms.*, Springer

[65] Bishop C M 1995 *Neural Networks for Pattern Recognition*, Oxford: Oxford University Press

[66] Bishop C M 2007 *Pattern Recognition and Machine Learning*, Springer-Verlag New York, USA

[67] Tankard Jr. W J 2001 *Applications of Computer Content Analysis*, Westport, CT: Ablex Publishing

[68] Mosteller F and Wallace D L 1963 J. Am. Stat. Assoc. **58** 302

[69] Witten I H and Frank E 2005 *Data Mining: Practical Machine Learning Tools and Techniques*, Morgan Kaufmann Publishers Inc., San Francisco, CA, USA

[70] Antiqueira L, Pardo T A S, Nunes M G V, Oliveira Jr. O N 2007 Inteligencia Artificial **11** 51–58

[71] Amancio D R, Comin C H, Casanova D, Travieso G, Bruno O M, Rodrigues F A and Costa L F (2014) PLoS ONE **9** (4): e94137

[72] Breiman L (1996) Mach. Learn. **24**: 123–140